We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

4

In Silico Analysis of Golgi Glycosyltransferases: A Case Study on the LARGE-Like Protein Family

Kuo-Yuan Hwa[1,2], Wan-Man Lin and Boopathi Subramani
Institute of Organic & Polymeric Materials
[1]Department of Molecular Science and Engineering
[2]Centre for Biomedical Industries
National Taipei University of Technology, Taipei, Taiwan, ROC

1. Introduction

Glycosylation is one of the major post-translational modification processes essential for expression and function of many proteins. It has been estimated that 1% of the open reading frames of a genome is dedicated to glycosylation. Many different enzymes are involved in glycosylation, such as glycosyltransferases and glycosidases.

Traditionally, glycosyltransferases are classified based on their enzymatic activities by Enzyme Commission (http://www.chem.qmul.ac.uk/iubmb/enzyme/). Based on the activated donor type, glycosyltransferases are named, for example glucosyltransferase, mannosyltransferase and *N*-acetylglucosaminyltransferases. However, classification of glycosyltransferases based on the biochemical evidence is a difficult task since most of the enzymes are membrane proteins. Reconstruction of enzymatic assay for the membrane proteins are intrinsically more difficult than soluble proteins. Thus the purification of membrane-bound glycosyltransferase is a difficult task. On the other hand, with the recent advancement of genome projects, DNA sequences of an organism are readily available. Furthermore, bioinformatics annotation tools are now commonly used by life science researchers to identify the putative function of a gene. Hence, new approaches based on *in silico* analysis for classifying glycosyltransferase have been used successfully. The best known database for classification of glycosyltransferase by *in silico* approach is the CAZy (Carbohydrate- Active enZymes) database (http://afmb.cnrs-mrs.fr/CAZy/) (Cantarel et al., 2009).

Glycosyltransferases are enzymes involved in synthesizing sugar moieties by transferring activated saccharide donors into various macro-molecules such as DNA, proteins, lipids and glycans. More than 100 glycosyltransferases are localized in the endoplasmic reticulum (ER) and Golgi apparatus and are involved in the glycan synthesis (Narimatsu, H., 2006). The structural studies on the ER and golgi glycosyltransferases has revealed several common domains and motifs present between them. The glycosyltransferases are grouped into functional subfamilies based on similarities of sequence, their enzyme characteristics, donor specificity, acceptor specificity and the specific donor and acceptor linkages (Ishida et al., 2005). The glycosyltransferase sequences comprise of 330-560 amino acids long and share the same type II transmembrane protein structure with four functional domains: a short

cytoplasmic domain, a targeting / membrane anchoring domain, a stem region and a catalytic domain (Fukuda et al., 1994). Mammals utilize only 9 sugar nucleotide donors for glycosyltransferases such as UDP-glucose, UDP-galactose, UDP-GlcNAc, UDP-GalNAc, UDP-xylose, UDP-glucuronic acid, GDP-mannose, GDP-fucose, and CMP-sialic acid. Other organisms have an extensive range of nucleotide sugar donors (Varki et al., 2008). Based on the structural studies, we have designed an intelligent platform for the LARGE protein, a golgi glycosyltransferase. The LARGE is a member of glycosyltransferase which has been studied in protein glycosylation (Fukuda & Hindsgaul, 2000). It was originally isolated from a region in chromosome 22 of the human genome which was frequently deleted in human meningiomas with alteration in glycosphingolipid composition. This led to a suggestion that the LARGE may have possible role in complex lipid glycosylation (Dumanski et al., 1987; Peyrard et al., 1999).

2. LARGE

LARGE is one of the largest genes present in the human genome and it is comprised of 660 kb of genomic DNA and contains 16 exons encoding a 756-amino-acid protein. It showed 98% amino acid identity to the mouse homologue and similar genomic organization. The expression of LARGE is ubiquitous but the highest levels of LARGE mRNA are present in heart, brain and skeletal muscle (Peyrard et al., 1999).

LARGE encodes a protein which has an N-terminal transmembrane anchor, coiled coil motif and two putative catalytic domains with a conserved DXD (Asp-any-Asp) motif typical of many glycosyltransferases that uses nucleoside diphosphate sugars as donors (Longman et al., 2003 & Peyrard et al., 1999). The proximal catalytic domain in the LARGE was most homologous to the bacterial glycosyltransferase family 8 (GT8 in CAZy database) members (Coutinho et al., 2003). The members of this family are mainly involved in the synthesis of bacterial outer membrane lipopolysaccharide. The distal domain resembled the human β1,3-N-acetytglucosaminyltransferase (iGnT), a member of GT49 family. The iGnT enzyme is required for the synthesis of the poly-N-acetyllactosamine backbone which is part of the erythrocyte *i* antigen (Sasaki et al., 1997). The presence of two catalytic domains in the LARGE is extremely unusual among the glycosyltransferase enzymes.

2.1 Functions of LARGE

2.1.1 Dystroglycan glycosylation

The Dystroglycan (DG) is an important constituent of the dystrophin-glycoprotein complex (DGC). This complex plays an essential role in the maintaining the stability of the muscle membrane and for the correct localization and/or ligand-binding activity, the glycosylation of some of these components are required (Durbeej et al., 1998). The DG comprises of two subunits, the extracellular α-DG and the transmembrane β-DG (Barresi, 2004). Various components present in the extracellular matrix including laminin (Smalheiser & Schwartz 1987), agrin (Gee et al., 1994), neurexin, (Sugita et al., 2001), and perlecan (Peng et al., 1998) interacts with α-DG. The carbohydrate moieties present in the α-DG are essential to bind with laminin and other ligands. The α-DG is modified by three different types of glycans such as: mucin type *O*-glycosylation, *O*-mannosylation, and *N*-glycosylation. The glycosylated α-DG is essential for the protein's ability to bind the laminin globular domain-containing proteins of the Extracellular Matrix (Kanagawa, 2005). LARGE is required for the generation of functional, properly glycosylated forms of α-DG (Barresi, 2004).

2.1.2 Human LARGE and α-Dystroglycan

The α-DG functional glycosylation by LARGE is likely to be involved in the generation of a glycan polymer which gives rise to the broad molecular weight range observed for α-DG detected by VIA4-1 and IIH6 antibodies. Both the human and mouse LARGE C-terminal glycosyltransferase domain is similar to β3GnT6, which adds GlcNAc to Gal to generate linear polylactosamine chains (Sasaki et al., 1997), the chain formed by LARGE might also be composed of GlcNAc and Glc.

In 1963, Myodystrophy, *myd*, was first described (Lane et al., 1976) as a recessive myopathy mapping to chromosome (Chr) 8, was identified as an intragenic deletion within the glycosyltransferase gene, LARGE. In Large^{myd} and enr mice, the hypoglycosylation of α-DG in DGC was due to the mutation in LARGE (Grewal et al., 2001). The α-DG function was restored in Large^{myd} skeletal muscle and ameliorates muscular dystrophy when LARGE gene was transferred, which indicated that adjustment in the glycosylation status of α-DG can improve the muscle phenotype.

The patients with clinical spectrum ranging from severe congenital muscular dystrophy (CMD), structural brain and eye abnormalities [Walker-Warburg syndrome (WWS), MIM 236670] to a relative mild form of limb-girdle muscular dystrophy (LGMD2I, MIM 607155) are linked to the abnormal O-linked glycosylation of α-DG (van Reeuwijk et al., 2005). A study made by Barresi R. *et al.* (2004) revealed the existence of dual and concentration dependent functions of LARGE. In physiological concentration, LARGE may be involved in regulating the α-DG O-mannosylation pathway. But when the LARGE is expressed by force, it may trigger some other alternative pathways for the *O*-glycosylation of α-DG which can generate a type of repeating polymer of variable lengths, such as glycosaminoglycan-like or core 1 or core 2 structures. This alternative glycan mimics the *O*-mannose glycan in its ability to bind α-DG ligands and can compensate for the defective tetrasaccharide. The functional LARGE protein is also required for neuronal migration during CNS development and it rescues α-DG in MEB fibroblasts and WWS cells (Barresi R. *et al.*, 2004).

2.1.3 LARGE in visual signal processing

The role of LARGE in proper visual signal processing was studied from the retina retinal pathology in Large^{myd} mice. The functional abnormalities of the retina was investigated by a sensitive tool called Electroretinogram (ERG). In Large^{myd} mice, the normal a-wave indicated that the mutant glycosyltransferase does not have any effect on its photoreceptor function.

But the alteration in b-wave may have resulted in downstream retinal circuitry with altered signal processing (Newman & Frishman, 1991). The DGC may also have a possible role in this aspect of the phenotype. The abnormal b-wave was responsible for the loss of retinal isoforms of dystrophin in humans and mice similar to the Large^{myd} mice.

2.2 LARGE homologues

A homologous gene to LARGE was identified and named as LARGE2. It is found to be involved in α-DG maturation as like LARGE, according to Fujimura et al., (2005). It is still not well understood whether these two proteins are compensatory or cooperative. The co-expression of LARGE and LARGE2 did not increase the maturation of α-DG in comparison with either one of them alone and it proved that for the maturation of α-DG, the function of LARGE2 is compensatory and not cooperative. Gene therapy for muscular dystrophy using the LARGE gene is a current topic of research (Barresi R. *et al.*, 2004; Braun, 2004). When compared to LARGE, LARGE2 gene may be more effective because it can glycosylate heavily than LARGE and it also prevents the harmful and immature α-DG production.

The closely related homologues of LARGE are found in the human genome, (glycosyltransferase-like 1B; GYLTL1B), mouse genome (Glylt1b; also called LARGE-Like or LargeL) and in some other vertebrate species (Grewal & Hewitt, 2002). The homologue gene is positioned on the chromosome 11p11.2 of the human genome and it encodes 721 amino acid protein which has 67% identity with LARGE, suggests that the two genes may have risen by gene duplication. Like LARGE, it is also predicted to have two catalytic domains, though it lacks the coiled-coiled motif present in the former protein. The hyperglycosylation of α-dystroglycan by the overexpression of GYLTL1B increased its ability to bind laminin and both the genes showed the same level of increase in laminin binding ability (Brockington, et al., 2005).

3. Bioinformatics workflow and platform design

Many public databases and bioinformatics tools have been developed and are currently available for use (Ding & Berleant, 2002). The primary goal of bioinformaticians is to develop reliable databases and effective analysis tools capable of handling bulk amount of biological data. But the objective of laboratory researchers is to study specific areas within the life sciences, which requires only a limited set of databases and analysis tools. Thus the existing free bioinformatics tools are sometimes too complicated for the biologists to choose. One solution is to have an expert team who are familiar with both bioinformatics databases and to know the needs of a research group in a particular field. The expert team will recommend a workflow by using selected bioinformatics tools and databanks and also helps the scientists with the complicated issue of tools and databases. Moreover, such a team could organize large number of heterogeneous sources of biological information into a specific, expertly annotated databank.

The team can also regularly and systematically update the information essential to help biologists overcome the problems of integrating and keeping up-to-date with heterogeneous biological information (Gerstein, 2000).

We have built a novel information management platform, LGTBase (Hyperlink).This composite knowledge management platform includes the "LARGE-like GlcNAc Transferase Database" by integrating specific public databases like CAZy database, and the workflow analysis combined the usage of specific, public & designed bioinformatics tools to identify the members of the LARGE-like protein family.

4. Tools and database selection

To analyze a novel protein family, biologists need to understand many different types of information. Moreover, the speed of discovery in biology has been expanding exponentially in recent years. So the biologists have to pick the right information available from the vast resources available. To overcome these obstacles, a bioinformatics workflow can be designed for analysing a specific protein family. In our study, a workflow was designed based on the structure and characteristics of LARGE protein as shown in Figure 1 (Hwa et al., 2007). The unknown DNA/protein sequences will be first identified as members of the known gene families by using the Basic Local Alignment Search Tool (BLAST). The *blastp* search tool is used to look for new LARGE-like proteins present in different organisms. The researchers who wish to use our platform can obtain the protein sequences either from the experimental data or through the *blastp* results. The search results were then analyzed with

the following tools. To begin with, the sequences are searched for the aspartate-any residue-aspartate (DXD) motif. The DXD motifs present in some glycosyltransferase families are essential for its enzyme activity.

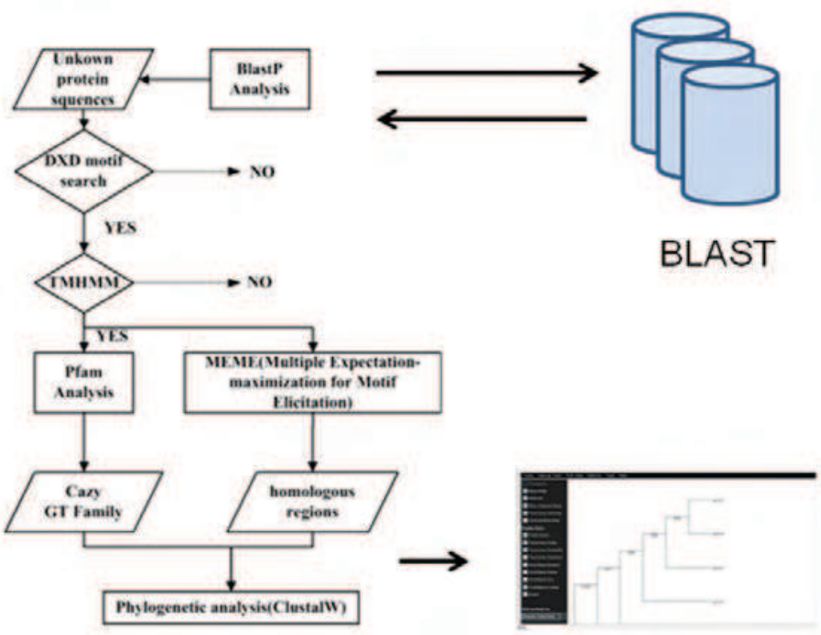


Fig. 1. Bioinformatics workflow of LGTBase.

The DXD motif prediction was then followed by the transmembrane domain prediction by using the TMHMM program (version 2.0; Center for Biological Sequence Analysis, Technical University of Denmark [http://www.cbs.dtu.dk/services/TMHMM-2.0/]). The transmembrane domain is a characteristic feature of the Golgi enzymes.

The sequence motifs are then identified by MEME (Multiple Expectation-maximization for Motif Elicitation) program (version 3.5.4; San Diego Supercomputer Center, UCSD [http://meme.sdsc.edu/meme/]).

This program finds the motif-homology between the target sequence and other known glycosyltransferases. In addition to all the above mentioned tools, the Pfam search (Sanger Institute [http://www.sanger.ac.uk/Software/Pfam/search.shtml]) can also be used to find the multiple sequence alignments and hidden Markov models in many existing protein domains and families. The Pfam results will indicate what kind of protein family the peptide belongs to. If it is a desired protein, investigators can then identify the evolutionary relationships by using phylogenetic analysis.

4.1 LARGE-like GlcNAc transferase database

The specific annotation entries used in the LGTBase are currently being used in a configuration that uses the information retrieved from several databases.

In CAZy database (Carbohydrate- Active enZymes) database ([http://afmb.cnrs-mrs.fr/CAZY/]), the glycosyltransferases are classified as families, clans, and folds based on their structural and sequence similarities, and also on their mechanistic investigation. The other databases used in this platform were listed in Table 1.

Database	Description	Website
EntrezGene	NCBI's repository for gene-specific information	http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene
GenBank	NIH genetic sequence database, an annotated collection of all publicly available DNA sequences	http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide
Dictybase	Database for model organism *Dictyostelium discoideum*	http://dictybase.org/
UniProtKB/Swiss-Prot	High-quality, manually annotated, non-redundant protein sequence database	http://www.uniprot.org/
InterPro	Database of protein families, domains and functional sites	http://www.ebi.ac.uk/interpro/
MGI	Database provides integrated genetic, genomic, and biological data of the laboratory mouse	http://www.informatics.jax.org/
Ensembl	It provides genome- annotation information	http://www.ensembl.org/index.html
HGMD	Human Gene Mutation Database (HGMD) provides comprehensive data on human inherited disease mutations	http://www.hgmd.cf.ac.uk/ac/index.php
UniGene	NCBI database of the transcriptome	http://www.ncbi.nlm.nih.gov/unigene
GeneWiki	The database transfers information on human genes to Wikipedia article	http://en.wikipedia.org/wiki/Gene_Wiki
TGDB	Database with information about the genes involved in cancers	http://www.tumor-gene.org/TGDB/tgdb.html
HUGE	The database provides the results of the Human cDNA project at the Kazusa DNA Research Institute	http://zearth.kazusa.or.jp/huge/
RGD	Database with collection of genetic and genomic information on the rat	http://rgd.mcw.edu/
OMIM	Database provides information on human genes and genetic disorders.	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
CGAP	Information of gene expression profiles of normal, precancer, and cancer cells.	http://cgap.nci.nih.gov/
PubMed	Database with 20 million citations for biomedical literature from medical journals, life science journals, related books.	http://www.ncbi.nlm.nih.gov/PubMed/
GO	Representation of gene and gene product attributes across all species	http://www.geneontology.org/

Table 1. The information sources of LARGE-like GlcNAc Transferase Database

All the information related to the LARGE-like protein family was retrieved from the different biological databases. In order to confirm that the information obtained was reliable, the data was scrutinized at two levels. First the information was selected from the above mentioned biological databases with customized programs (using the *perl* compatible regular expressions). Then the obtained information was annotated and validated by experts in glycobiology and bioinformatics.

The annotated data in the LGTBase database was divided into nine categories (Figure 2). The first category is related to genomic location, displays the chromosome, the cytogenetic band and the map location of the gene. The second is related to aliases and descriptions, displays synonyms and aliases for the relevant gene, and descriptions of its function, cellular localization and effect on phenotype. The third category on proteins provides annotated information about the proteins encoded by the relevant genes. The fourth is about protein domains and families, provides annotated information about protein domains and families and the fifth on protein function which provides annotated information about gene function. The sixth category is related to pathways and interactions, provides links to pathways and interactions followed by the seventh on disorders and mutations which draws its information from OMIM and UniProt. The eighth category is on expression in specific tissues, shows that the tissue expression values are available for a given gene. The last category is about research articles, lists the references related to the proteins which are studied. In addition, the investigator can also use DNA or protein sequences to assemble the dataset for the analysis using this workflow.

Fig. 2. The contents of LGTBase database

4.2 LARGE-like GlcNAc transferase workflow

4.2.1 Reference sequences search

The unknown DNA/protein sequences are identified as members of the known gene families using the Basic Local Alignment Search Tool (BLAST). BlastP is one of the BLAST programs and it searches protein databases using a protein query. We used BlastP to look for new LARGE-like proteins from different species and gathered the protein sequences of

LARGE like GlcNAc Transferases and built a protein database of 'LARGE-like protein'. This database would assist in search for more reference sequences of LARGE-like protein.

4.2.2 DXD motif search

In several glycosyltransferase families, the DXD motif is essential for the enzymatic activity (Busch et al. 1998). So we first searched for aspartate-any residue-aspartate (DXD) motif, commonly found in glycosyltransferase. Therefore, the 'DXD Motif Search' tool was designed. The input protein sequences are loaded or pasted in this tool and the results indicate the presence or absence of DXD motif.

4.2.3 Transmembrane helices search

The LARGE protein is a member of the N-acetylglucosaminyltransferase family. The presence of transmembrane domain is a characteristics feature of this family. TMHMM program is used to predict the transmembrane helices based on the hidden Markov model. The prediction gives the most probable location and orientation of transmembrane helices in the sequence. TMHMM can predict the location of transmembrane alpha helices and the location of intervening loop regions. This program also predicts the location of the loops that are present between the helices either inside or outside of the cell or organelle. The program is designed based on a 20 amino acids long alpha helix which contains hydrophobic amino acids that can span through a cell membrane.

4.2.4 MEME analysis

A motif is a sequence pattern that occurs repeatedly in a group of related protein or DNA sequences. MEME (Multiple Expectation-maximization for Motif Elicitation) represents motifs as position-dependent letter-probability matrices which describe the probability of each possible letter at each position in the pattern. The program can search for homologous sequences among the input protein sequences.

4.2.5 Protein families search

The Pfam HMM search was used to identify the protein family to which the input protein sequences belong. The Pfam database contains the information about most of the protein domains and families. The results from the Pfam HMM search will show the relation of input protein sequences with the existing protein families and domains.

4.2.6 Phylogenetic analysis

The phylogenetic analysis was performed to find any significant evolutionary relationship between the new protein sequences and the LARGE protein family and to support our previous findings. ClustalW, a multiple alignment program which aligns two or more sequences to determine any significant consensus sequences between them (Thompson et al., 1994). This approach can also be used for searching patterns in the sequence. The phylogenetic tree was constructed by using PHYLIP program (v.3.6.9) and viewed by Treeview software (v.1.6.6). In GlcNAc-transferase phylogenetic analysis, once the multiple alignment of all GlcNAc-transferase has been done, it can be used to construct the phylogenetic tree. About 25 protein sequences were identified as the LARGE-like protein family. By using the neighbor joining distance method, the phylogenetic tree showed that these proteins can be divided into 6 groups (Figure 3). The evolutionary history inferred

from phylogenetic analysis is usually depicted as branching, tree-like diagrams which represents an estimated pedigree of the inherited relationships among the protein sequences from different species. These evolutionary relationships can be viewed either as Cladograms (Chenna et al., 2003) or Phylograms (Radomski & Slonimski, 2001).

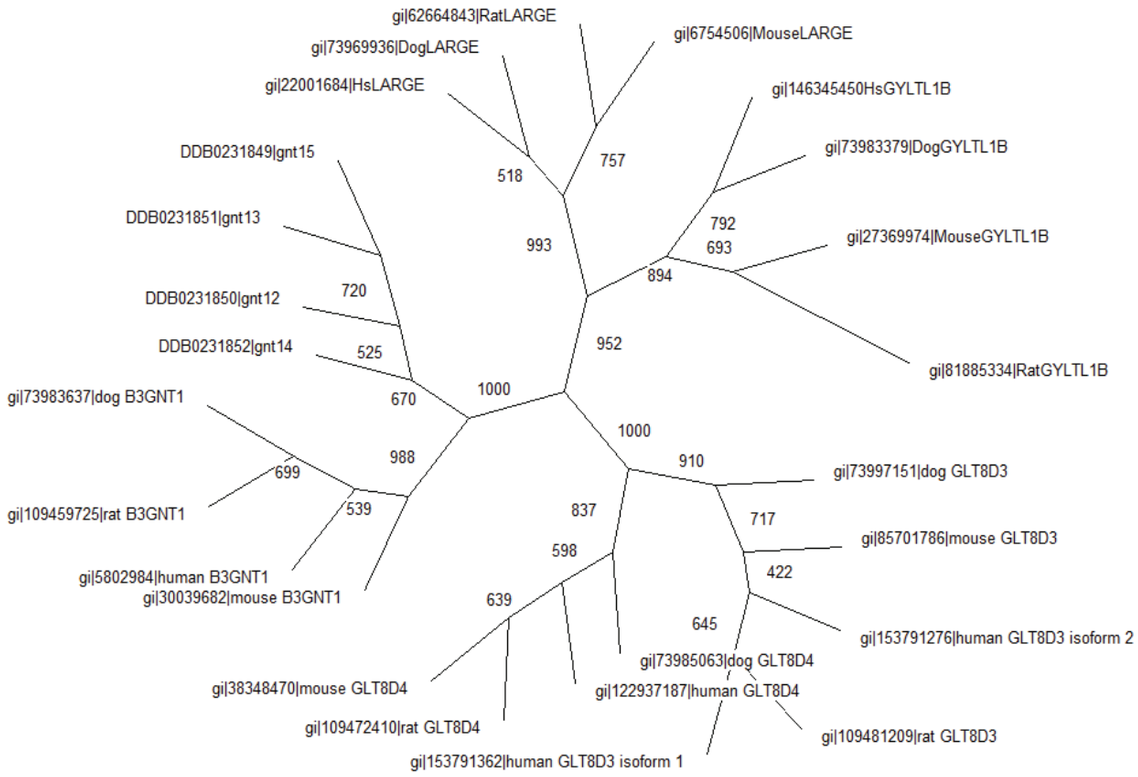


Fig. 3. Phylogenetic tree of LARGE-like Protein Family

4.3 Organization of the LGTBase platform

The data obtained from the analyses were stored in a MySQL relational database and the web interface was built by using PHP and CGI/Java scripts. According to the characteristics of LARGE-like GlcNAc transferase proteins, the workflow was designed and developed by using Java language and several open source bioinformatics programs. Tools with different languages, C, *perl*, java were integrated by using Java language (Figure 4). Adjustable parameters of the tools were reserved to fulfill the needs in future.

5. Application with LARGE protein family

A protein sequence (fasta format) can be entered into the BlastP assistant interface, enabling the other known proteins with similar sequences to be identified (Figure 5). The investigator can select all the resulting sequences or use only some of them. The data can then be transferred to the DXD analysis page (Figure 6). The rationale behind choosing the DXD analysis was since they are represented in many families of glycosyltransferases and it will be easy to narrow down the analysis of putative protein sequences to particular protein families or domains. There were many online tools available for the identification and

characterization of unknown protein sequences. So depending upon the target protein of study, one can pick the tools to characterize it.

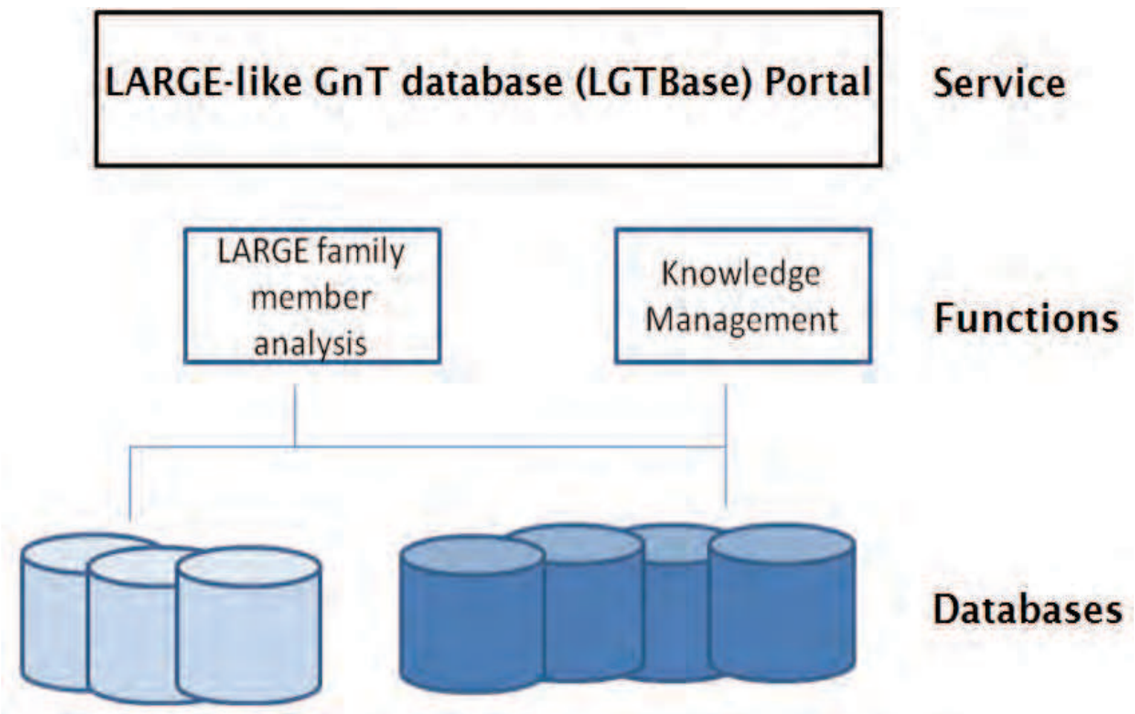


Fig. 4. Database selected for construction of the knowledge management platform

The sequences are analyzed with the DXD motif search tool (Figure 6), which selects those sequences containing the DXD motif for the TMHMM analysis. The transmembrane helices can be predicted with TMHMM analysis (Figure 7). The transmembrane domains are predicted by the hydrophobic nature of the proteins and mainly used to identify the cellular location of the proteins. Similar to transmembrane domain prediction, there were several other domains that can be predicted based on the protein's characters like hydrophobic, hydrophilic etc., The dataset containing DXD motifs and transmembrane helices are then selected for MEME (Figure 8) and Pfam analysis (Figure 9). Some sequence motifs occur repeatedly in the data set and are conjectured to have a biological significance are predicted by MEME analysis. This application plays a significant role in characterization of the putative protein sequences after the initial studies with the DXD motif, transmembrane domain, and other tools. This tool can be used for all kind of protein sequences since its prediction is based on the pattern of sequences present in the study. The protein sequences in the dataset can be identified to the known protein families by Pfam analysis. The pfam classification can also be used for almost all the putative protein sequences because of its large collection of protein domain families represented by multiple sequence alignments and Hidden Markov Models. After the MEME and Pfam analysis were done, ClustalW and Phylip programs were used for Phylogenetic Analysis (Figure 9) to see the evolutionary relationship among the data sets (Figure 10). Finally, these results can be used to design experiments to be performed in the laboratory.

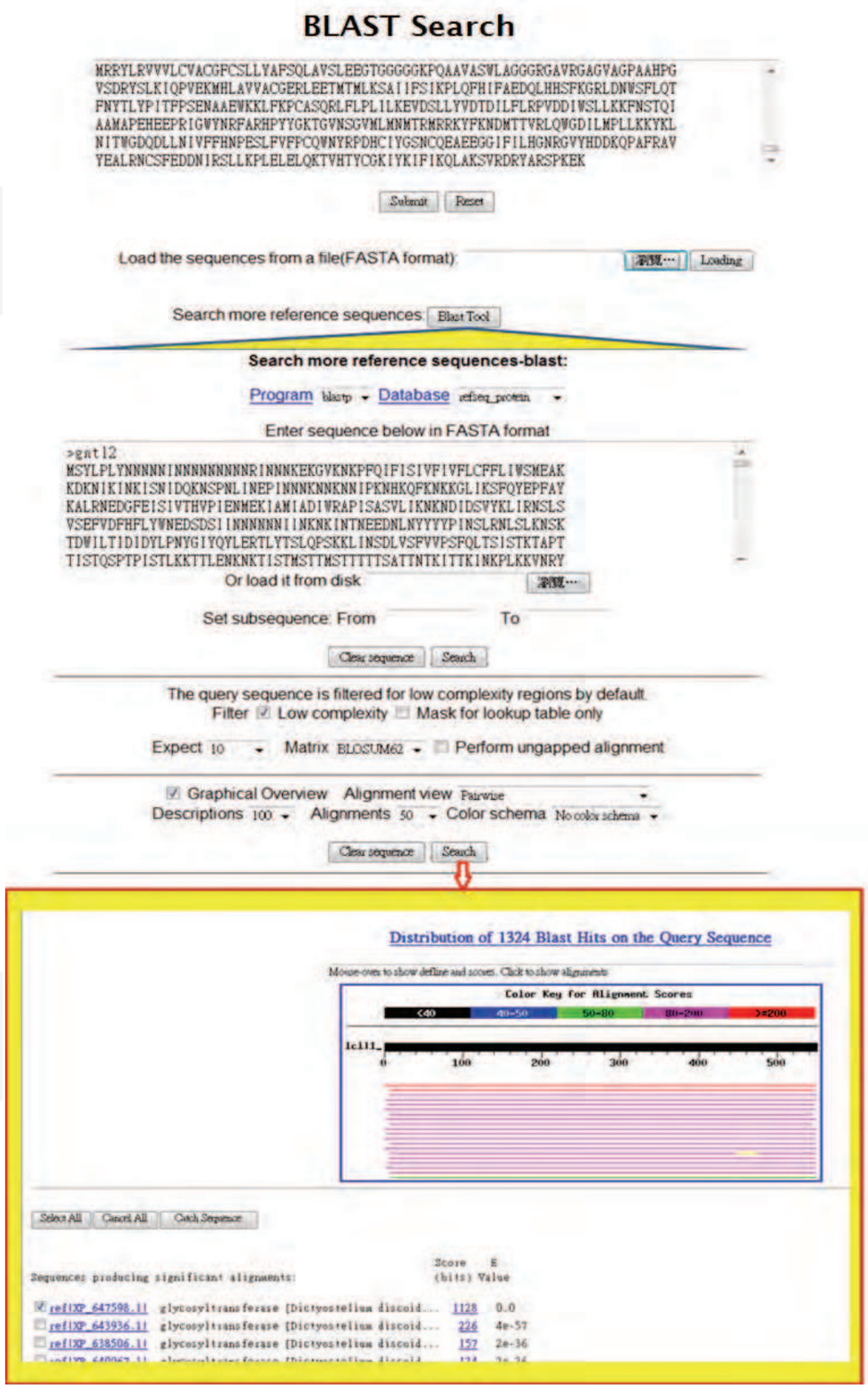


Fig. 5. The BlastP tool of the LGTBase platform to find similar sequences to LARGE

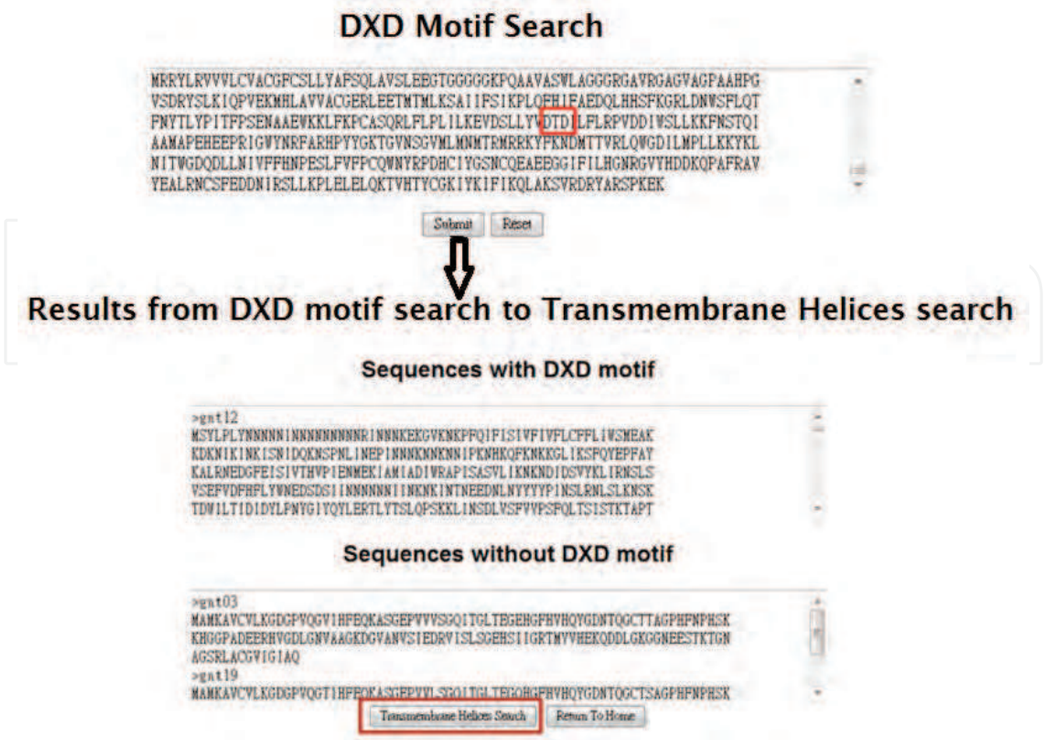


Fig. 6. DXD motif search tool of the LGTBase platform for DXD motif prediction

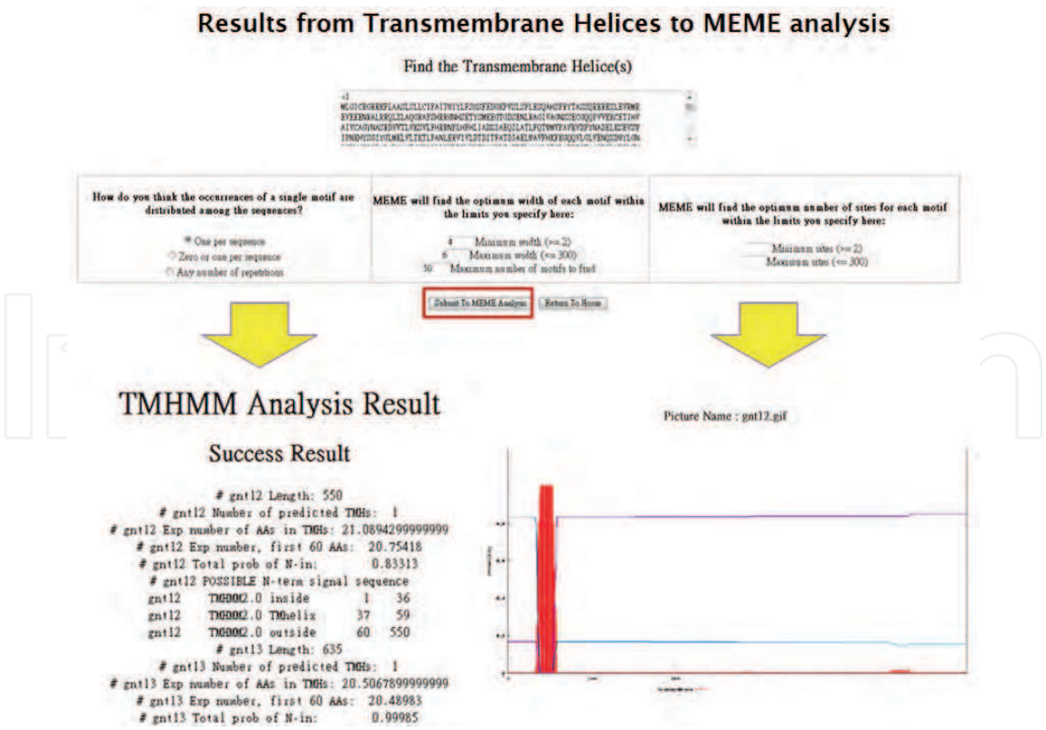


Fig. 7. TMHMM analysis tool of the LGTBase platform for Transmembrane domain prediction

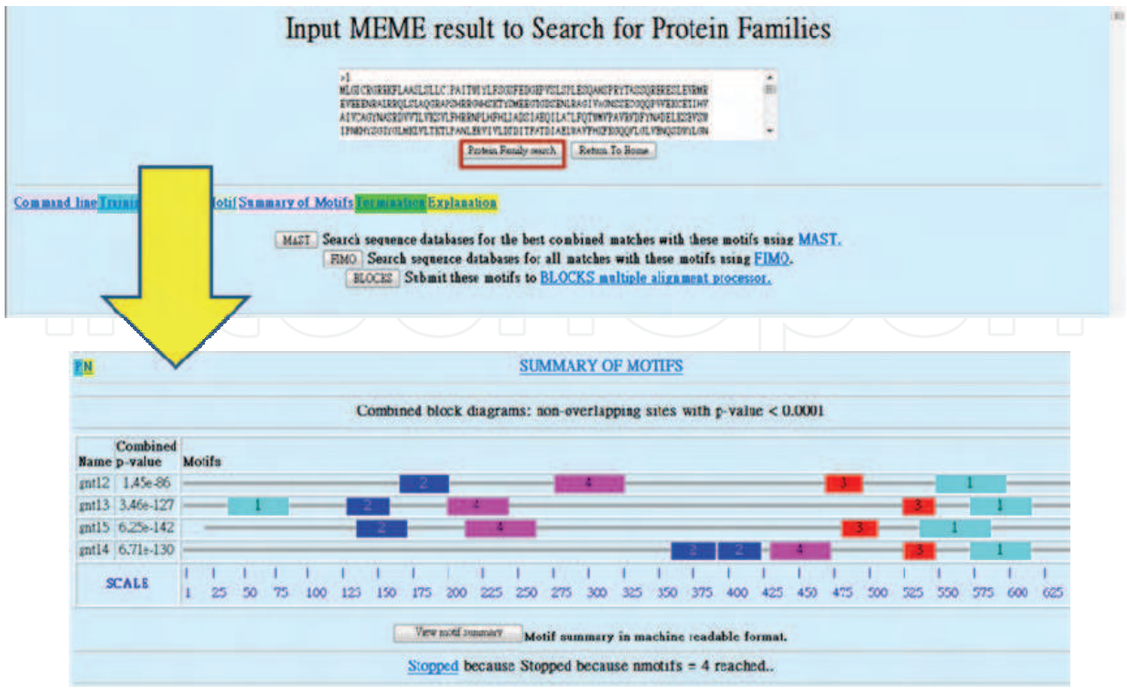
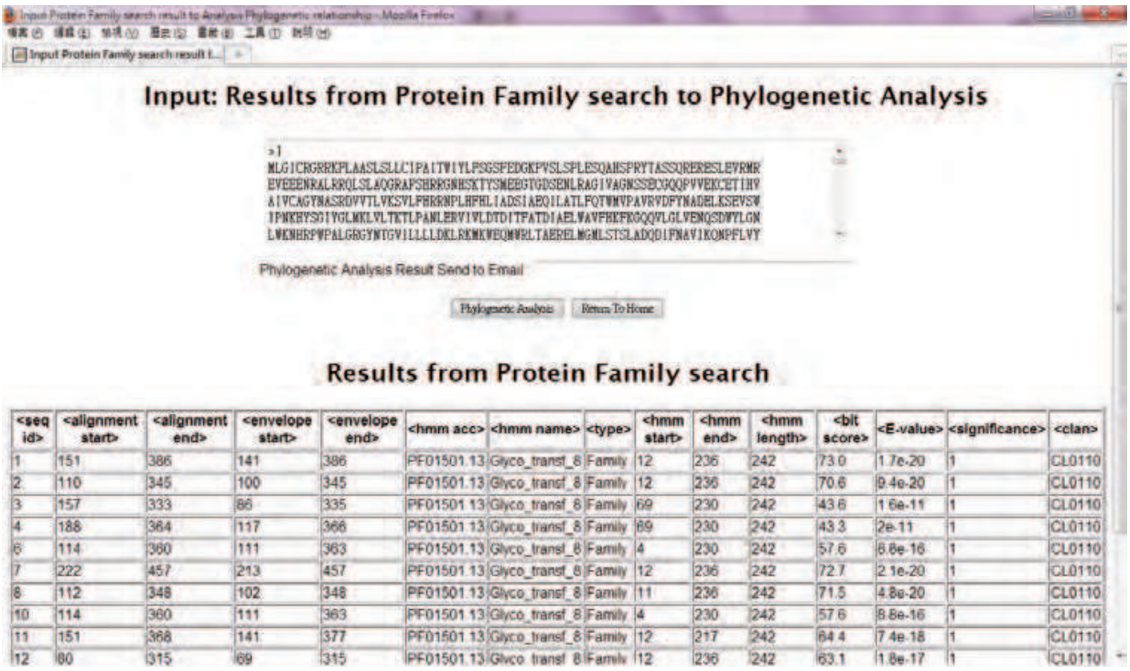


Fig. 8. MEME analysis tool of the LGTBase platform to predict the sequence motifs



<seq id>	<alignment start>	<alignment end>	<envelope start>	<envelope end>	<hmm acc>	<hmm name>	<type>	<hmm start>	<hmm end>	<hmm length>	<bit score>	<E-value>	<significance>	<clan>
1	151	386	141	386	PF01501.13	Glyco_transf_8	Family	12	236	242	73.0	1.7e-20	1	CL0110
2	110	345	100	345	PF01501.13	Glyco_transf_8	Family	12	236	242	70.6	9.4e-20	1	CL0110
3	157	333	86	335	PF01501.13	Glyco_transf_8	Family	69	230	242	43.6	1.6e-11	1	CL0110
4	188	364	117	366	PF01501.13	Glyco_transf_8	Family	69	230	242	43.3	2e-11	1	CL0110
6	114	360	111	363	PF01501.13	Glyco_transf_8	Family	4	230	242	57.6	8.6e-16	1	CL0110
7	222	457	213	457	PF01501.13	Glyco_transf_8	Family	12	236	242	72.7	2.1e-20	1	CL0110
8	112	348	102	348	PF01501.13	Glyco_transf_8	Family	11	236	242	71.5	4.8e-20	1	CL0110
10	114	360	111	363	PF01501.13	Glyco_transf_8	Family	4	230	242	57.6	8.8e-16	1	CL0110
11	151	368	141	377	PF01501.13	Glyco_transf_8	Family	12	217	242	64.4	7.4e-18	1	CL0110
12	80	315	69	315	PF01501.13	Glyco_transf_8	Family	12	236	242	63.1	1.8e-17	1	CL0110

Fig. 9. Pfam analysis tool of the LGTBase platform to identify the known protein family of the target protein which is studied

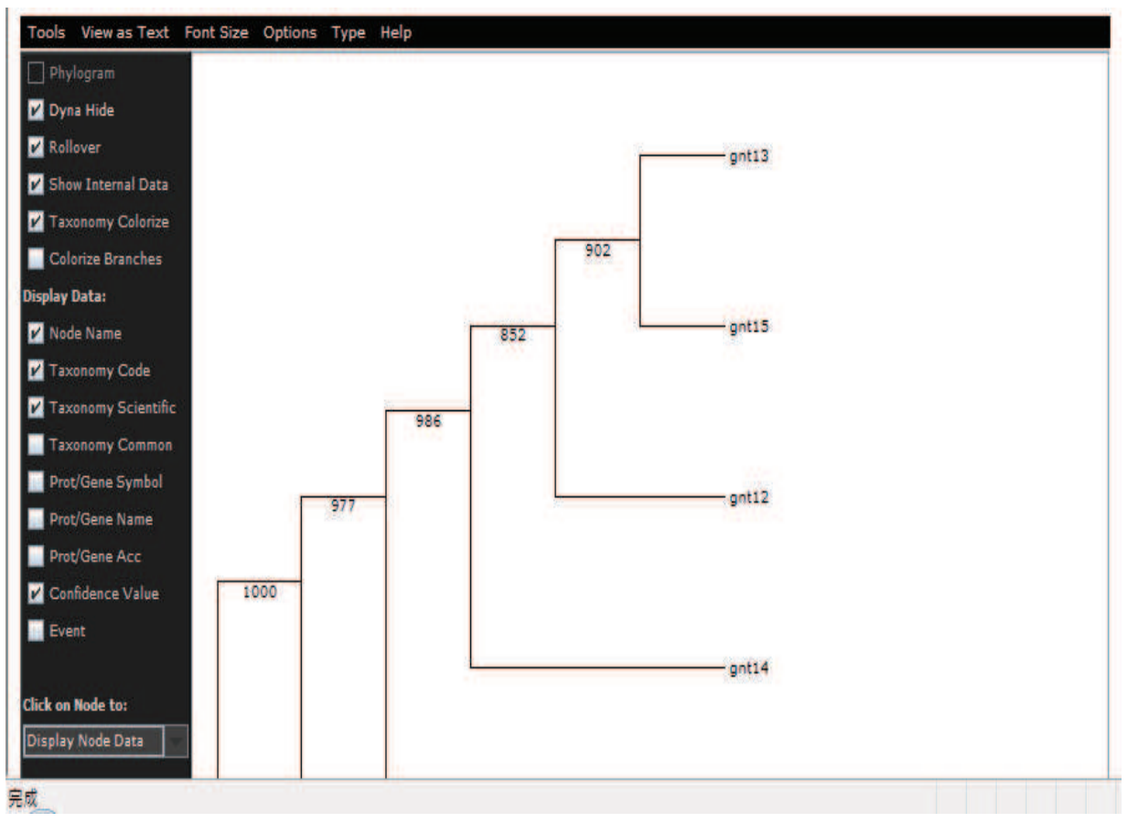


Fig. 10. Phylogenetic analysis tool of the LGTBase platform to study the evolutionary relationship of the target protein

6. Future direction

We have described how to construct a computational platform to analyze the LARGE protein family. Since the platform was built based on several commonly shared protein domains and motifs, it can also be modified for analyzing other golgi glycosyltransferases. Furthermore, the phylogenetic analysis (Figure 3) revealed that LARGE protein family is related to β-1,3-*N*-acetylglucosaminyltransferase 1 (β3GnT). β3GnT (EC 2.4.1.149) is a group of enzymes belong to glycosyltransferases family. Some β3GnT enzymes catalyze the transfer of GlcNAc from UDP-GlcNAc to Gal in the Galβ1-4 Glc(NAc) structure with β-1,3 linkage. These enzymes were grouped into GT family 31, 49 in the CAZy database. The enzyme uses 2 substrates namely, UDP-N-acetyl-D-glucosamine and D-galactosyl-β-1,4-N-acetyl-D-glucosaminyl-R and the products are formed as UDP, N-acetyl-β-D- glucosaminyl-

β-1,3 -D-galactosamine. These enzymes participate in the formation of keratan sulfate, glycosphingolipid biosynthesis, neo-lacto series and N-linked glycans. There are currently 9 members known from the β3GnT family.

The β3GnT1 (iGnT) was the first enzyme to be isolated when cDNA of a human β-1,3-N-acetylglucosaminyltransferase essential for poly-N-acetyllactosamine synthesis was studied (Zhou et al., 1999). The poly-N-acetyllactosamine synthesized by iGnT provides critical backbone structure for the addition of functional oligosaccharides such as Sialyl Lewis X. It has been reported recently that β3GnT1 is involved in attenuating prostate cancer cell locomotion by regulating the synthesis of laminin-binding glycans on α-DG (Bao et al., 2009). Since there are several common shared domains similar to the LARGE protein, the new platform for β3GnT protein family can be constructed based on the original platform. Apart from β3GnT1, β3GnT2 enzyme is responsible for elongation of poly-lactosamine chains. This enzyme was isolated based on structural similarity with the β3GalT family. Studies showed that on a panel of invasive and noninvasive fresh transitional cell carcinomas (TCCs) showed strong down regulation of β3GnT2 in the invasive lesions, suggesting that a decline in the expression levels of some members of the glycosyltransferase (Gromova et al., 2001).

The β3GnT3 and β3GnT4 enzymes were subsequently isolated based on the structural similarity with β3GalT family. β3GnT3 is a type II transmembrane protein and contains a signal anchor that is not cleaved. It prefers the substrates of lacto-N-tetraose and lacto-N-neotetraose, and it is also involved in the biosynthesis of poly-N-acetyllactosamine chains and the biosynthesis of the backbone structure of dimeric sialyl Lewis A. It plays dominant role in the L-selectin ligand biosynthesis, lymphocyte homing and lymphocyte trafficking. The β3GnT3 enzyme is highly expressed in the non-invasive colon cancer cells. β3GnT4 is involved in the biosynthesis of poly-N-acetyllactosamine chains and prefers lacto-N-neotetraose as the substrate. It is a type II transmembrane protein and it is expressed more in bladder cancer cells (Shiraishi et al., 2001). β3GnT5 is responsible for lactosyltriaosylceramide synthesis, an essential component of lacto/neolacto series glycolipids (Togayachi et al., 2001). The expression of the HNK-1 and Lewis x antigens on the lacto/neo-lacto-series of glycolipids is developmentally and tissue-specifically regulated by β3GnT5. The overexpression of β3GnT5 in human gastric carcinoma cell lines led to increased sialyl-Lewis X expression and increased *H.pylori* adhesion (Marcos et al., 2008).

The β3GnT6 synthesizes the core 3 O-glycan structure and speculates that this enzyme plays an important role in the synthesis and function of mucin O-glycan in the digestive organs. In addition, the expression of β3GnT6 was markedly down regulated in gastric and colorectal carcinomas (Iwai et al., 2005). Expression of β3GnT7 has been reported to be down-regulated upon malignant transformation (Kataoka et al., 2002). Elongation of the carbohydrate backbone of keratan sulfate proteoglycan is catalyzed by β3GnT7 and β1,4-galactosyltransferase 4 (Hayatsu et al., 2008). β3GnT7 can transfer GlcNAc to Gal to synthesize a polylactosamine chain with each enzyme differing in its acceptor molecule preference. The polylactosamine and related structures plays crucial role in cell-cell interaction, cell-extracellular matrix interaction, immune response and determining metastatic capacity. The β3GnT8 enzyme extends a polylactosamine chain specifically on a tetraantennary N-glycans. β3GnT8 transfers GlcNAc to the non-reducing terminus of the

Galβ1-4GlcNAc of tetra antennary *N*-glycan *in vitro*. Intriguingly, β3GnT8 is significantly upregulated in colon cancer tissues than in normal tissue (Ishida et al., 2005). The co-transfection of β3GnT8 and β3GnT2 resulted in synergistic enhancement of the activity of the polylactosamine synthesis. This indicates that these two enzymes interact and complement each other's function in the cell. As a summary, the members of the β3GnT protein family are important in human cancer biology.

Our initial motif analysis showed that there are 3 important functional domains predicted are commonly found among the β3GnT enzymes. The first motif is a structural motif necessary for maintaining the protein fold. The second, DXD motif represented in many glycosyltransferases is involved in the binding of the nucleotide-sugar donor substrate, both directly and indirectly through coordination of metal ions such as magnesium or manganese in the active site. A glycine-rich loop is the third motif found at the bottom of the active site cleft. This loop is likely to play a role in the recognition of both the GlcNAc portion of the donor and the substrate. Since the three common domains of β3GnT are similar to the LARGE protein family, it is feasible to modify the current LARGE platform to analyze other golgi glycosyltransferases such as β3GnT.

7. References

Bao, X., Kobayashi, M., Hatakeyama, S., Angata, K., Gullberg, D., Nakayama, J., Fukuda, M.N. & Fukuda, M. (2009). Tumor suppressor function of laminin-binding α-dystroglycan requires a distinct β-3-N-acetylglucosaminyltransferase. *Proceedings of the National Academy of Sciences USA,* Vol.106, No.29, (July 2009), pp. 12109-12114

Barresi, R., Michele, D.E., Kanagawa, M., Harper, H.A., Dovico, S.A., Satz, J.S., Moore, S.A., Zhang, W., Schachter, H., Dumanski, J.P., Cohn, R.D., Nishino, I. & Campbell, K.P. (2004). LARGE can functionally bypass alpha-dystroglycan glycosylation defects in distinct congenital muscular dystrophies. *Nature Medicine,* Vol.10, No.7, (July 2004), pp. 696-703.

Braun, S. (2004). Naked plasmid DNA for the treatment of muscular dystrophy. *Current Opinion in Molecular Therapeutics,* Vol.6, (October 2004), pp. 499-505.

Brockington, M., Torelli, S., Prandini, P., Boito, C., Dolatshad, N.F., Longman, C., Brown, S.C., Muntoni, F. (2005). Localization and functional analysis of the LARGE family of glycosyltransferases: significance for muscular dystrophy. *Human Molecular Genetics*, Vol.14, No.5, (March 2005), pp. 657-665.

Busch, C., Hofmann, F., Selzer, J., Munro, S., Jeckel, D. & Aktories, K. (1998). A common motif of eukaryotic glycosyltransferases is essential for the enzyme activity of large clostridial cytotoxins. *Journal of Biological Chemistry,* Vol.273, No.31, (July 1998), pp.19566-19572.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research, Vol.* 37, (January 2009), pp. D233-238

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J. & Higgins, D.G. Thompson JD (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* Vol.31, No.13, (July 2003), pp. 3497-3500.

Coutinho, P.M., Deleury, E., Davies, G.J. & Henrissat, B. (2003). An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology*, Vol.328, (April 2003), pp. 307-317.

Ding, J., Berleant, D., Nettleton, D. & Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases? *Pacific Symposium on Biocomputing*, Vol.7, pp. 326-337.

Dumanski, J.P., Carlbom, E., Collins, V.P., Nordenskjold, M. (1987). Deletion mapping of a locus on human chromosome 22 involved in the oncogenesis of meningioma. *Proceedings of the National Academy of Sciences USA*, Vol.84, (December 1987), pp. 9275-9279.

Durbeej, M., Henry, M.D. & Campbell, K.P. (1998). Dystroglycan in development and disease. *Current Opinions in Cell Biology* Vol. 10, (October 1998), pp. 594-601.

Fujimura, K., Sawaki, H., Sakai, T., Hiruma, T., Nakanishi, N., Sato, T., Ohkura, T., Narimatsu, H. (2005). LARGE2 facilitates the maturation of *α*-dystroglycan more effectively than LARGE. *Biochemical and Biophysical Research Communications,* Vol.329, No.3, (April 2005), pp. 1162-1171

Fukuda, M., Hindsgaul, O., Hames, B.D. & Glover, D.M. (1994). In *Molecular Glycobiology*, Oxford Univ. Press, Oxford.

Fukuda, M., & Hindsgaul, O. (2000). *Molecular and Cellular Glycobiology* (2nd ed.), Oxford Univ. Press, Oxford.

Gee, S.H., Montanaro, F., Lindenbaum, M.H., and Carbonetto, S. (1994). Dystroglycan-α, a dystrophin-associated glycoprotein, is a functional agrin receptor. *Cell,* Vol.77, (June 1994), pp. 675-686.

Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nature Structural Biology,* Vol.7, (November 2000), Suppl: 960-963.

Grewal, K., Holzfeind, P.J., Bittner, R.E. & Hewitt, J.E., (2001). Mutant glycosyltransferase and altered glycosylation of alpha-dystroglycan in the myodystrophy mouse. *Nature Genetics*, Vol.28, (June 2001), pp.151-154.

Grewal, P.K. & Hewitt, J.E. (2002). Mutation of Large, which encodes a putative glycosyltransferase, in an animal model of muscular dystrophy. *Biochimica et Biophysica Acta*, Vol.1573, (December 2002), pp. 216-224.

Gromova, I., Gromov, P. & Celis J.E. (2001). A Novel Member of the Glycosyltransferase Family, β3GnT2, highly down regulated in invasive human bladder Transitional Cell Carcinomas. *Molecular Carcinogenesis,* Vol. 32, No. 2, (October 2001), pp. 61-72

Hayatsu, N., Ogasawara, S., Kaneko, M.K., Kato, Y. & Narimatsu, H. (2008). Expression of highly sulfated keratan sulfate synthesized in human glioblastoma cells. *Biochemical and Biophysical Research Communications,* Vol. 368, No. 2, (April 2008), pp. 217-222

Hwa, K.Y., Pang, T.L. & Chen, M.Y. (2007). Classification of LARGE-like GlcNAc-Transferases of *Dictyostelium discoideum* by Phylogenetic Analysis. *Frontiers in the Convergence of Bioscience and Information Technologies*, pp. 289-293.

Ishida, H., Togayachi, A., Sakai, T., Iwai, T., Hiruma, T., Sato, T., Okubo, R., Inaba, N., Kudo, T., Gotoh, M., Shoda, J., Tanaka, N., & Narimatsu, H. A novel beta1,3-N-acetylglucosaminyltransferase (beta3Gn-T8), which synthesizes poly-N-acetyllactosamine, is dramatically upregulated in colon cancer. *FEBS Letters*. (January 2005), Vol. 579, No.1, pp. 71-78.

Ishida, H., Togayachi, A., Sakai, T., Iwai, T., Hiruma, T., Sato, T., Okubo, R., Inaba, N., Kudo, T., Gotoh, M., Shoda, J., Tanaka, N. & Narimatsu, H. (2005). A novel beta1,3-N-acetylglucosaminyltransferase (beta3Gn-T8), which synthesizes poly-N-acetyllactosamine, is dramatically upregulated in colon cancer. *FEBS Letters*, Vol.579, No.1, (January 2005), pp. 71-8.

Iwai, T., Kudo, T., Kawamoto, R., Kubota, T., Togayachi, A., Hiruma, T., Okada, T., Kawamoto, T., Morozumi, K. & Narimatsu, H. (2005). Core 3 synthase is down-regulated in colon carcinoma and profoundly suppresses the metastatic potential of carcinoma cells. *Proceedings of the National Academy of Sciences USA,* Vol.102, No.12, (March 2005), pp. 4572-4577

Kanagawa, M., Michele, D.E., Satz, J.S., Barresi, R., Kusano, H., Sasaki, T., Timpl, R., Henry, M. D., and Campbell, K.P. (2005). Disruption of Perlecan Binding and Matrix Assembly by Post-Translational or Genetic Disruption of Dystroglycan Function. *FEBS Letters,* Vol.579, No.21, (August 2005), pp. 4792-4796.

Kataoka, K. & Huh, N.H. (2002). A novel β1,3-N-acetylglucosaminyltransferase involved in invasion of cancer cells as assayed *in vitro*. *Biochemical and Biophysical Research Communications,* Vol. 294, No.4, (June 2002), pp. 843-848

Lane, P.W., Beamer, T.C. & Myers, D.D. (1976). Myodystrophy, a new myopathy on chromosome 8 of the mouse. *Journal of Heredity*, Vol. 67, No.3 (May-June 1976), pp. 135-138.

Longman, C., Brockington, M., Torelli, S., Jimenez-Mallebrera, C., Kennedy, C., Khalil, N., Feng, L., Saran, R.K., Voit, T., Merlini, L., Sewry, C.A., Brown, S.C. & Muntoni F. (2003). Mutations in the human LARGE gene cause MDC1D, a novel form of congenital muscular dystrophy with severe mental retardation and abnormal glycosylation of alpha dystroglycan. *Human Molecular Genetics*, Vol.12, No.21, (November 2003), pp. 2853-2861.

Marcos, N.T., Magalhães, A., Ferreira, B., Oliveira, M.J., Carvalho, A.S., Mendes, N., Gilmartin, T., Head, S.R., Figueiredo, C., David, L., Santos-Silva, F. & Reis, C.A. (2008). *Helicobacter pylori* induces β3GnT5 in human gastric cell lines, modulating expression of the SabA ligand Sialyl-Lewis X. *Journal of Clinical Investigation*, Vol. 118, No. 6, (June 2008), pp.2325-2336

Narimatsu, H. (2006). Human glycogene cloning: focus on beta 3-glycosyltransferase and beta 4-glycosyltransferase families. *Current Opinions in Structural Biology*. Vol.16, No.5, (October 2006), pp. 567-575.

Newman, E.A. & Frishman, L.J. (1991). The b-wave. In Arden, G.B. (ed.*), Principles and Practice of Clinical Electrophysiology of Vision*, Mosby-Year Book, St Louis, MO.

Peng, H.B., Ali, A.A., Daggett, D.F., Rauvala, H., Hassell, J.R., & Smalheiser, N.R. (1998). The relationship between perlecan and dystroglycan and its implication in the

formation of the neuromuscular junction. *Cell Adhesion and Communication,* Vol.5, No.6, (September 1998), pp. 475-489

Peyrard, M., Seroussi, E., Sandberg-Nordqvist, A.C., Xie, Y.G., Han, F.Y.,Fransson, I., Collins, J., Dunham, I., Kost-Alimova, M., Imreh, S.,Dumanski, J.P., (1999). The human LARGE gene from 22q12.3-q13.1 is a new, distinct member of the glycosyltransferase gene family. *Proceedings of the National Academy of Sciences USA*, Vol.96, No.2, (January 1999), pp. 589-603.

Radomski, J.P. & Slonimski, P.P. (2001). Genomic style of proteins: concepts, methods and analyses of ribosomal proteins from 16 microbial species. *FEMS Microbiol Reviews,* Vol.25, No.4, (August 2001), pp. 425-435.

Sasaki, K., Kurata-Miura, K., Ujita, M., Angata, K., Nakagawa, S., Sekine, S., Nishi, T. & Fukuda, M. (1997). Expression cloning of cDNA encoding a human beta-1,3-N-acetylglucosaminyl transferase that is essential for poly-N-acetyllactosamine synthesis. *Proceedings of the National Academy of Sciences USA,* Vol.94, No.26, (December 1997), pp. 14294-14299.

Shiraishi, N., Natsume, A., Togayachi, A., Endo, T., Akashima, T., Yamada, Y., Imai, N., Nakagawa, S., Koizumi, S., Sekine, S., Narimatsu, H. & Sasaki K. (2001). Identification and characterization of 3 novel β1,3-N-Acetylglucosaminyltransferases. Structurally Related to the β1,3-Galactosyltransferase family. *The Journal of Biological Chemistry,* Vol. 276, No.5, (February 2001), pp. 3498-3507

Smalheiser, N. R., and Schwartz, N. B. (1987) Cranin: a laminin-binding protein of cell membranes. *Proceedings of the National Academy of Sciences USA,* Vol.84, No.18, (September 1987), pp. 6457-6461.

Sugita, S., Saito, F., Tang, J., Satz, J., Campbell, K., & Sudhof, T.C. (2001). A stoichiometric complex of neurexins and dystroglycan in brain. *Journal of Cell Biology,* Vol.154, No.2, (July 2001), pp. 435-445

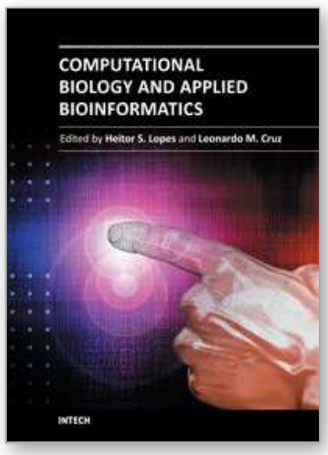
Thompson, J.D., Higgins, D.G. & Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* Vol.22, No.22, (November 1994), pp. 4673-4680.

Togayachi, A., Akashima, T., Ookubo, R., Kudo, T., Nishihara, S., Iwasaki, H., Natsume, A., Mio, H. Inokuchi J. and T. Irimura *et al.*, Molecular cloning and characterization of UDP-GlcNAc: Lactosylceramide β1,3-*N*-acetylglucosaminyltransferase (β3Gn-T5), an essential enzyme for the expression of HNK-1 and Lewis X epitopes on glycolipids, *Journal of Biological Chemistry,* Vol. 276, No.5, (March 2001), pp. 22032–22040

van Reeuwijk, J., Brunner, H.G., van Bokhoven, H. (2005). Glyc-O-genetics of Walker-Warburg syndrome. *Clinical Genetics,* Vol. 67, No.4, (April 2005), pp. 281-289.

Varki, A., Cummings, R.D., Esko, J.D., Freeze, H.H., Stanley, P., Bertozzi, C.R., Hart, G.W. & Etzler, M.E. (2008). *Essentials of Glycobiology*, (2nd ed.) Plainview (NY): Cold Spring Harbor Laboratory Press

Zhou, D., Dinter, A., Gutiérrez Gallego, R., Kamerling, J.P., Vliegenthart, J.F., Berger, E.G. & Hennet, T. (1999). A β-1,3-N-acetylglucosaminyltransferase with poly-N-acetyllactosamine synthase activity is structurally related to β-1,3-galactosyltransferases. *Proceedings of the National Academy of Sciences USA,* Vol. 96, No. 2, (January 1999), pp. 406-411



Computational Biology and Applied Bioinformatics

Edited by Prof. Heitor Lopes

ISBN 978-953-307-629-4
Hard cover, 442 pages
Publisher InTech
Published online 02, September, 2011
Published in print edition September, 2011

Nowadays it is difficult to imagine an area of knowledge that can continue developing without the use of computers and informatics. It is not different with biology, that has seen an unpredictable growth in recent decades, with the rise of a new discipline, bioinformatics, bringing together molecular biology, biotechnology and information technology. More recently, the development of high throughput techniques, such as microarray, mass spectrometry and DNA sequencing, has increased the need of computational support to collect, store, retrieve, analyze, and correlate huge data sets of complex information. On the other hand, the growth of the computational power for processing and storage has also increased the necessity for deeper knowledge in the field. The development of bioinformatics has allowed now the emergence of systems biology, the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of a living being. This book presents some theoretical issues, reviews, and a variety of bioinformatics applications. For better understanding, the chapters were grouped in two parts. In Part I, the chapters are more oriented towards literature review and theoretical issues. Part II consists of application-oriented chapters that report case studies in which a specific biological problem is treated with bioinformatics tools.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen